IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DIABETES DETECTION USING MACHINE LEARNING

Mr. CH. Paparao[1], K Kishore Kumar[2], P Naresh Kumar[3], I Prashnathi[4], A Mahesh Babu[5]

[1]Associate.prof, CSE Dept, GVR&S CET, Guntur, Andhra Pradesh, India

[2,3,4,5]Resarch Scholar, CSE Dept, GVR&S CET, Guntur, Andhra Pradesh, India

Abstract: The primary goal of this project is to use various machine learning approaches to predict the potential presence of diabetes at an early stage, with a focus on females. Making the appropriate lifestyle adjustments at the right time will help prevent diabetes and all the problems connected with it. Early detection of diabetes can considerably minimize the disease's progression and lower the chance of catastrophic consequences including heart and kidney ailments. Therefore, a device that can help physicians identify this fatal illness earlier and halt its course is desperately needed. Finally, using the support vector machine classifier model as a base, this model generated an accuracy of 78%.

Index Terms - Machine learning, diabetes, mellitus, Pima Indians dataset, etc.,

I. INTRODUCTION

Diabetes is a disease where glucose, or blood sugar, is not metabolized by the body which increases the glucose rate to alarmingly high levels. Normally, a hormone called insulin helps control the amount of glucose in one's bloodstream, people with diabetes either don't produce insulin (type 1 diabetes) or don't respond to insulin the way they should (type 2 diabetes). Approximately, 90% of all diagnosed cases of diabetes are that of type 2. According to, the number of people living with diabetes more than tripled between 1990 and 2010, and the number of new cases doubled every year. Why are the numbers rising so fast? Obesity is believed to account for 8085% of the risk of developing type 2 diabetes and the World Health Organization (WHO) studies have shown that worldwide obesity has nearly tripled since 1975, this leads to the belief that the escalating rates of obesity and type 2 diabetes are directly linked to each other. One of the reasons for the global rise in obesity is that people are eating more high-calorie, high-fat foods and are less physically active because new technological advancements provide entertainment, education, communication, and all types of purchases right on the spot. This model focuses on the early detection of type 2 diabetes because it is more common. To carry out the training and testing of the machine learning model, the Pima Indians dataset from the National Institute of Diabetes and Digestive and Kidney Diseases was used. All patients in this dataset are females who are at least 21 years old and with Pima Indian Heritage, Pima Indians in the United States have the world's highest recorded prevalence and incidence of type 2 diabetes which is why the study was conducted on this specific group to generate this dataset. It consists of 8 medical predictor variables (attributes), and a single target, outcome. The outcome is the variable that specifies if a patient has been diagnosed with type 2 diabetes or not. The dataset contains 768 instances. The remaining part of the paper is organized as follows: Section II includes previous work that addressed the same problem. Section III introduces the complex details of the used dataset, the process of preparing the data and making it suitable for a machine learning model, and the machine learning algorithms used. Moreover, the results of every technique used and the associated accuracy of it are presented in Section IV. At last, a conclusion is outlined in section V.Machine learning has been used successfully for the prediction of many outcomes, ranging from the likelihood of being admitted to a university, predicting what books you might like based on your history, or even predicting who tweeted a certain tweet. However, a more relevant case is the use of machine learning in the detection of heart disease using majority ensemble methods. In addition, multiple algorithms and models have been trained in the field of diabetes detection, and several methods have been used to perform data preprocessing. In, a dataset consisting of 178131 instances has been used to train a model that reached an accuracy of 80.8%. The model used the Random Forest Classifier method when using all 14 physical examination features including age, pulse rate, height, weight, fasting glucose, etc. A model using the Pima Indians Dataset3 used the k-nearest neighbor (KNN) algorithm and tried different k-values ranging from 1 to 100 to reach a maximum receiver operating characteristic accuracy of 74% when k was set to 0. Moreover, the paper is a study to build an effective prediction model to identify Canadian patients at risk of having Diabetes Mellitus based on patient demographic data and the laboratory results during their visits to medical facilities, it has been trained on a dataset that contains 13309 Canadian patients with their ages ranging between 18 and 90 years. The Gradient Boosting Machine (GBM) technique performed best according to the evaluation of the area under the receiver operating characteristic curve (AROC), the AROC for this model is 84.7% with a sensitivity of 71.6%.

Table 1: PIMA Indian Dataset Attributes Description

Attributes	Range	Description
Pregnancies	0-17	Number of times pregnant
Glucose	0-199	Plasma glucose concentration 2 hours in an oral glucose tolerance test
Blood Pressure	0-122	Diastolic blood pressure (mm Hg)
BMI	0-67.1	Body mass index = (weight in kg/(height inm)^2)
Skin Thickness	0-99	Triceps skin fold thickness (mm)
Diabetesn Pedigree Function	0.078-2.42	A function that scores the likelihood of diabetes based on family history
Age	21-81	Age in years
Insulin	0-846	2-Hour serum insulin (mu U/ml)
Outcome	0-1	Class variable, diagnoses classes: 0 = healthy,1 =diagnosed with diabetes

Finally, the research used multiple techniques on different datasets. The algorithms used included Naïve Bayesian, Random Forest (RF), and KNN and used evaluation techniques like K-fold Cross-Validation. The highest accuracy achieved on the Pima Indian dataset (which was used as an example of a numeric-only dataset in the research) was 64.47% using the k-fold cross-validation.

II. LITERATURE SURVEY

Nour Abdulhadi, Amjed Almousa, the main objective of this research is to predict the possible presence of diabetes -specifically in females- at an early stage using different machine learning techniques. Early detection of diabetes can significantly prevent the progression of the disease and reduce the risk of serious complications such as heart and kidney diseases, making the proper lifestyle changes at the right time can help avoid diabetes and all the illnesses associated with it. So, there is a crucial need for a tool that can better assist doctors in detecting this deadly disease at an early stage and consequently stop its progression. Finally, this model produced an accuracy of 82% based on the random forest classifier model.

KM Jyoti RANI, Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled to 592 million. Diabetes is a disease caused due to the increased level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is one of the leading causes of blindness, kidney failure, amputations, heart failure, and stroke. When we eat, our body turns food into sugars or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, allowing the glucose to enter, and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with how machines learn from experience. This project aims to develop a system that can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. Algorithms like K nearest neighbour, Logistic Regression, Random Forest, Support vector machine, and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting diabetes. Decision Tree has the most accuracy 99%.

Boshra Farajollahi, Maysam Mehmannavaz, Hafez Mehrjoo, Fateme Moghbeli, Diabetes is a disease associated with high levels of glucose in the blood. Diabetes causes many kinds of complications, which also leads to a high rate of repeated admission of patients with diabetes. This study aims to diagnose Diabetes with machine learning techniques. Adaboost has the most accuracy 83%.

Md. Ashraf Uddin, Md. Manowa rulIslam, Md. Alamin Talukder, MdAl Amin Hossain, Arn isha Akhter, Sunil Aryal, Maisha Muntaha , Diabetes is a chronic disease characterized by the inability of the pancreas to produce enough insulin or the body's inability to use insulin efficiently. This disease is becoming increasingly prevalent worldwide and can result in severe complications such as blindness, kidney failure, and stroke. Early detection of diabetes can potentially save millions of lives globally, making it a crucial focus of research. In this study, we propose a machine learning model to aid in predicting diabetes. The model comprises several machine learning methods: Linear Regression (LnR), Logistic Regression (LR), k-nearest neighbor (KNN), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). Before feeding the preprocessed data into the machine learning model for evaluation, we conducted several pre-processing steps, such as removing null values, standardizing data using normalization, and labeling data using the label encoding process. Imbalanced datasets can adversely affect the accuracy of machine learning algorithms, and we address this problem by balancing the datasets using the Synthetic Minority Oversampling Technique (SMOTE) method. We assessed the model's performance on two datasets and found that the random forest algorithm produced optimal results, with 97% accuracy on the diabetes dataset 2019 and 80% accuracy on the Pima Indian dataset.

Rian Budi Lukmantoa , Suharjitoa Ariadi , Nugrohoa , Habibullah Akbara , The number of patients that were infected by Diabetes Mellitus (DM) has reached 415 million patients in 2015 and by 2040 this number is expected to increase to approximately 642 million patients. Large amount of medical data of DM patients is available and it provides significant advantage for researchers to fight against DM. The main objective of this research is to leverage F-Score Feature Selection and Fuzzy Support Vector Machine

in classifying and detecting DM. Feature selection is used to identify the valuable features in dataset. SVM is then used to train the dataset to generate the fuzzy rules and Fuzzy inference process is finally used to classify the output. The aforementioned methodology is applied to the Pima Indian Diabetes (PID) dataset. The results show a promising accuracy of 89.02% in predicting patients with DM. Additionally, the approach taken provides an optimized count of Fuzzy rules while still maintaining sufficient accuracy.

N.Sneha and Tarun Gangil, Diabetes is a chronic disease or group of metabolic disease where a person sufers from an extended level of blood glucose in the body, which is either the insulin production is inadequate, or because the body's cells do not respond properly to insulin. The constant hyperglycemia of diabetes is related to long-haul harm, brokenness, and failure of various organs, particularly the eyes, kidneys, nerves, heart, and veins. The objective of this research is to make use of significant features, design a prediction algorithm using Machine learning and find the optimal classifier to give the closest result comparing to clinical outcomes. The proposed method aims to focus on selecting the attributes that ail in early detection of Diabetes Miletus using Predictive analysis. The result shows the decision tree algorithm and the Random Forest has the highest specificity of 98.20% and 98.00%, respectively holds best for the analysis of diabetic data. Naïve Bayesian outcome states the best accuracy of 82.30%.

Roshan Birjais, Ashish Kumar Mourya, Ritu Chauhan,Harleen Kaur, Machine learning is a subset of Artificial Intelligence when combined with Data Mining techniques plays a promising role in the feld of prediction. We live in an era where data generation is exponential with time but if the generated data is not put to work or not converted to knowledge data, its generation is of no use. Similarly, in Healthcare also, data availability is high, so is the need to extract the information from it for better prognosis, diagnosis, treatment, drug development, and overall healthcare. In this research, we have tried to focus more on diagnosis of Diabetes disease, which is one of the fastest growing chronic diseases all over the world as declared by World Health Organization in the year 2014. We have also tried to show the diferent techniques like Gradient Boosting, Logistic Regression and Naive Bayes, which can be used for the diagnosis of diabetes disease with attained accuracy as 86% for the Gradient Boosting, 79% for Logistic Regression and 77% for Naive Bayes.

Dr. AEEVWIEKPAEFE, ABDULKADIR Nafisat, Diabetes Mellitus (DM) which refers to a metabolic disorder that occurs when the level of blood sugar in the body is considered high, which could be a resulting effect of inadequate availability of insulin in the body. It is a chronic disease which may lead to myriads of complications in the body system. Statistics by the World Health Organization (WHO) in 2013, indicated that DM was the cause of death of over 1.5 million people around the world and in 2016, 8.5% of adults within age seventeen (17) and above were reported to be diabetic and diabetic patients have continued to increase in recent years. It is therefore very glaring that these alarming figures calls for very urgent and effective attention. There has been a recent proliferate increase in studies relating to machine learning in the healthcare sector, hence the motivation for this research work. The research was based on the prevalence of diabetes amongst the masses of Kaduna metropolis using some selected hospitals as a case study after which a predictive model was designed for diabetes, using some selected supervised learning algorithms like Decision tree algorithm, K- Nearest Neighbour algorithm and Artificial Neural Networks on a dataset gotten from 44 Army Reference Hospital and Yusuf Danstoho Memorial Hospital Kaduna which constitutes of nine (9) attributes that was considered. The results indicated that ANN produced the highest accuracy with 97.40% followed by decision tree algorithm with 96.10% accuracy then K-NN algorithm with 88.31% accuracy. This result was further validated using fifty (50) dataset out of which forty-eight results were rightly predicted.

Hakim El Massari, Zineb Sabouri, Sajida Mhammedi and Noreddine Gherabi, Diabetes is one of the chronic diseases, which is increasing from year to year. The problems begin when diabetes is not detected at an early phase and diagnosed properly at the appropriate time. Different machine learning techniques, as well as ontology-based ML techniques, have recently played an important role in medical science by developing an automated system that can detect diabetes patients. This paper provides a comparative study and review of the most popular machine learning techniques and ontology-based Machine Learning classification. Various types of classification algorithms were considered namely: SVM, KNN, ANN, Naive Bayes, Logistic regression, and Decision Tree. The results are evaluated based on performance metrics like Recall, Accuracy, Precision, and F-Measure that are derived from the confusion matrix. The experimental results showed that the best accuracy goes for ontology classifiers and SVM.

Omer Faruk Akmese, The rate of diabetes is rapidly increasing worldwide. Early detection of diabetes can help prevent or delay the onset of diabetes by initiating lifestyle changes and taking appropriate preventive measures. Prediabetes and type 2 diabetes have proved to be early detection problems. There is a need for easy, rapid, and accurate diagnostic tools for the early diagnosis of diabetes in this context. Machine learning algorithms can help diagnose diseases early. Numerous studies are being conducted to improve the speed, performance, reliability, and accuracy of diagnosing with these methods for a particular disease. This study aims to predict whether a patient has diabetes based on diagnostic measurements in a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. Eight different variables belonging to the patients were selected as the input variable, and it was estimated whether the patient had diabetes or not. Of the 768 records examined, 500 (65.1%) were healthy, and 268 (34.9%) had diabetes. Ten different machine learning algorithms have been applied to predict diabetic status. The most successful method was the Random Forest algorithm with 90.1% accuracy. Accuracy percentages of other algorithms are also between 89% and 81%. This study describes a highly accurate machine learning prediction tool for finding patients with diabetes.

III.EXPERIMENTAL SETUP

The main purpose of this paper is to build a model that predicts diabetes at an early stage using the previously mentioned dataset. It is a real-world dataset taken from a specific group in a specific area as previously mentioned. Part of the data will be used to train the model, and the other to test it making it able to adapt to new unknown data to predict the outcome.

Dataset Attribute Information Each of the 767 instances in the dataset has 9 attributes, one of them being the target variable. A description of each attribute is present in Table 1. To get a further insight into the data, correlation values were calculated to know how much an attribute affects the target attribute (Outcome) or if other attributes are affected by it. Correlation values were calculated using the Pearson (product-moment) correlation coefficient equation. It computes the ratio of the covariance of both features to the product of their standard deviations consequently finding the measure of the linear relationship between those two features. Correlation values are shown in Table 2.

Table 2: Correlation with Outcome (Target)

Attribute	Correlation Value
Pregnancies	0.22
Glucose	0.47
Blood Pressure	0.07
BMI	0.29
Skin Thickness	0.07
Diabetes Pedigree Function	0.17
Age	0.24
Insulin	0.13

The heat map of the calculated correlation values is shown in Figure:1 below.

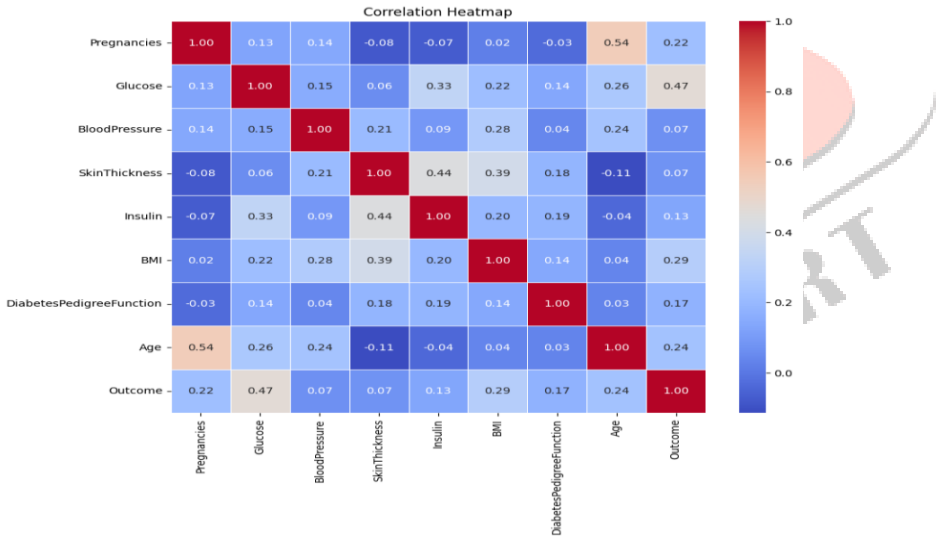
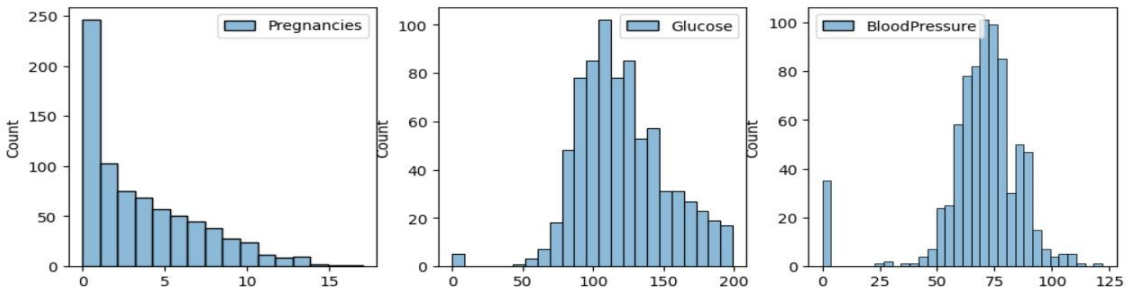


Figure: 1 Heat map to show the correlation between features



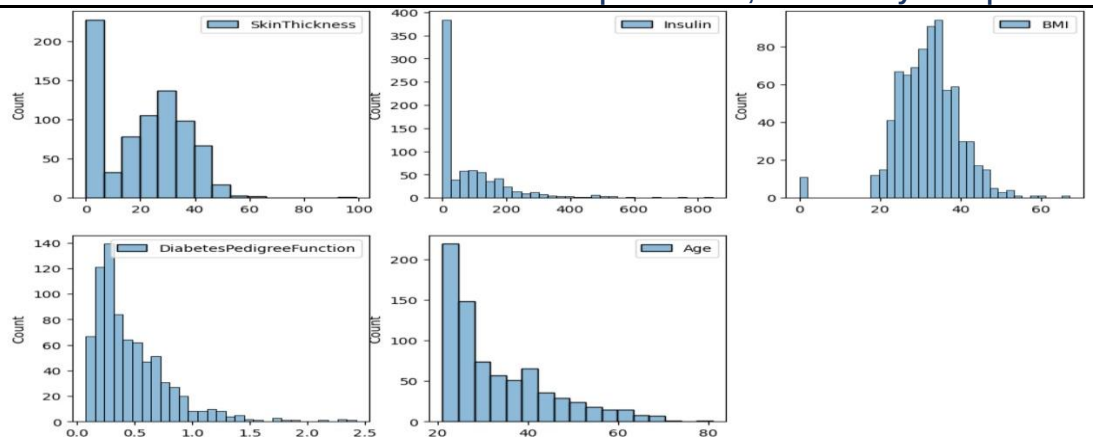


Figure:2 Histograms of the different attributes

It can be observed that glucose has the highest positive correlation with the outcome variable, followed by Diabetes Pedigree Function. Moreover, histograms were generated to have a better visual interpretation of the data, shown in Figure 2. In addition to the better visualization histograms provide, the figures can make it easier to detect possible outliers that may negatively affect the proposed model.

Data Preprocessing The quality of the data used to train the model significantly affects the results, especially when exposed to new data. Real-world data can contain errors or missing values, as well as outliers. Preprocessing of data helps minimize the effect of such errors, increasing the success rate of the project at hand. In the Pima Indian Dataset, multiple values were missing from a couple of instances. Having zero blood pressure, for example, does not make any sense. Since the number of instances present (768) was quite low, instead of dropping instances with zeros, the values were filled with the mean. Please note that Figure 1 was generated after the missing values were filled in. Also, the dataset had different scales, so it had to be standardized. Skipping this step could lead to the contribution of a feature more than the other to the target, whereas when the range of all features is normalized each feature contributes approximately proportionately to the final decision. The dataset was standardized using a standard scaler.

IV. MACHINE LEARNING ALGORITHM

After analyzing the data and filling in all the missing values in attributes such as blood pressure, skin thickness, and BMI, the data was split into two parts: test set and training set. The training set will be used to test the model, while the test set will be used to validate the ability of the model to generalize to new data. The classifier models that have been tested are:

4.1 Logistic Regression Classifier The first model that was used is the Logistic Regression Classifier, it is similar to the linear regression model that computes a weighted sum of the input features, but instead of outputting the result as the Linear Regression does, it outputs the logistic of the result. It models the chance of a certain outcome based on individual characteristics.

4.2 K-Nearest Neighbors classifier K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems.

4.3 Support Vector Machine (SVM) Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three

.

4.4 Decision Tree Classifier Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

4.5 Random Forest Classifier This method is one of the simplest and most diverse algorithms used for both classification and regression tasks, it uses multiple individual decision trees to operate as a single one. Each tree classifies the class to which an instance belongs, and the class with the highest votes is the predicted class.

V.RESULTS AND DISCUSSION

we show the performance of machine learning classification techniques diabetes classification. For this, we analyze various popular classification techniques that include the logistic regression, decision tree, support vector machine, knn, random forest. Support vector machine area archived best accuracy to compare another algorithm.

Table 5: Results Of Accuracy

Model Name	Accuracy
Logistic Regression	75%
K-Nearest Neighbors classifier	72%
Support Vector Machine (SVM)	78%
Decision Tree Classifier	70%
Random Forest Classifier	74%

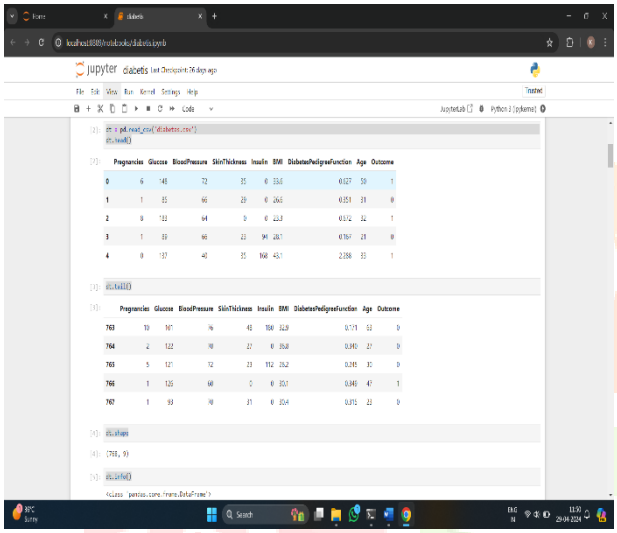


Figure5.1: Head and Tail

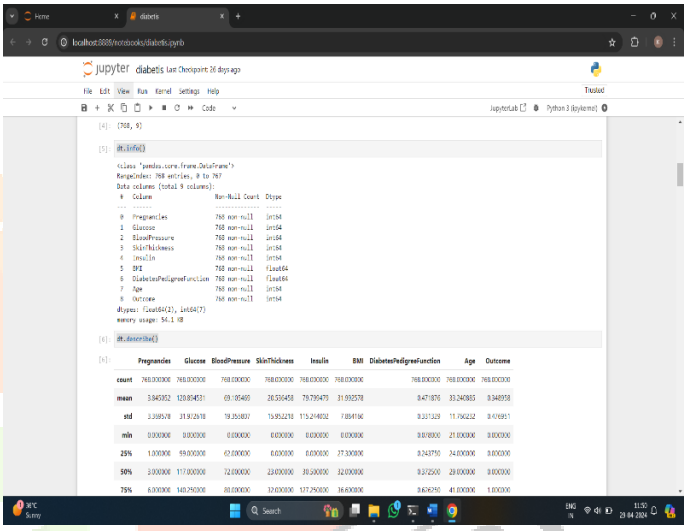


Figure5.2: Information

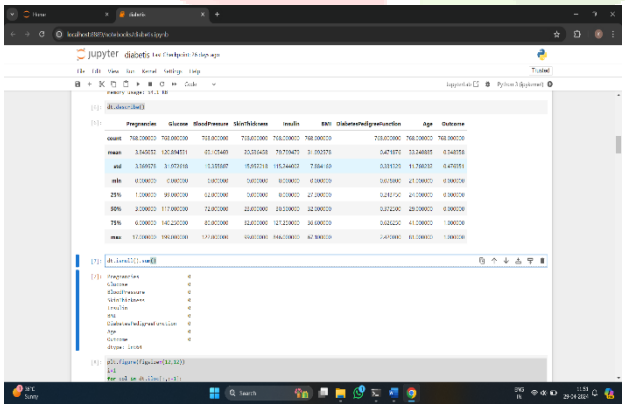


Figure 5.3: Is null

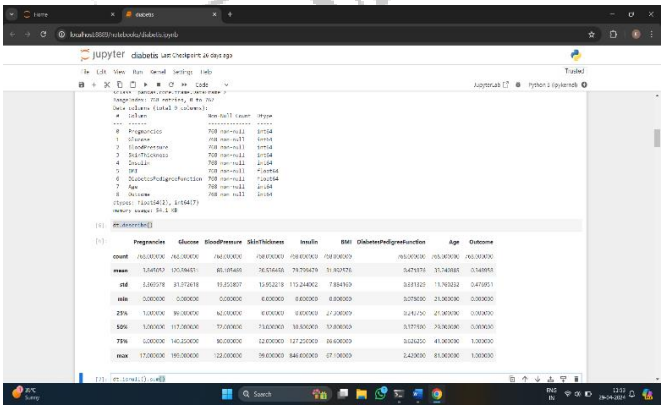


Figure 5.4: Describe

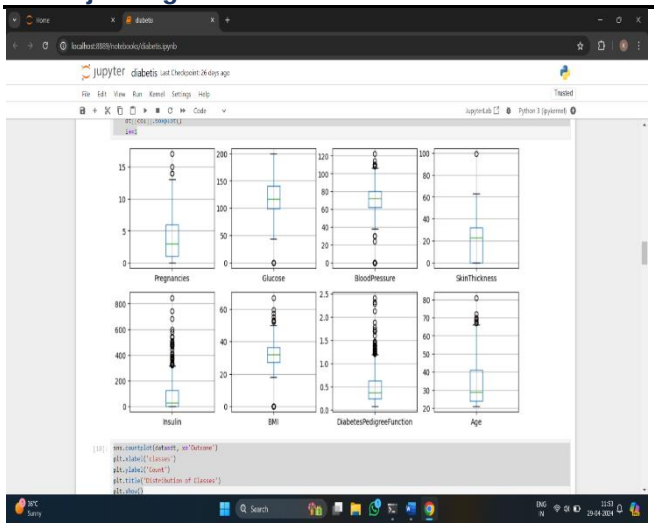


Figure 5.5: Box Plot

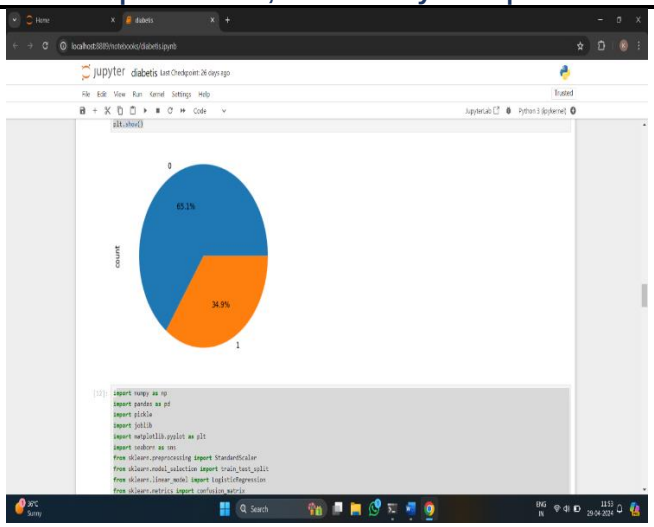


Figure 5.6: Pie Chart

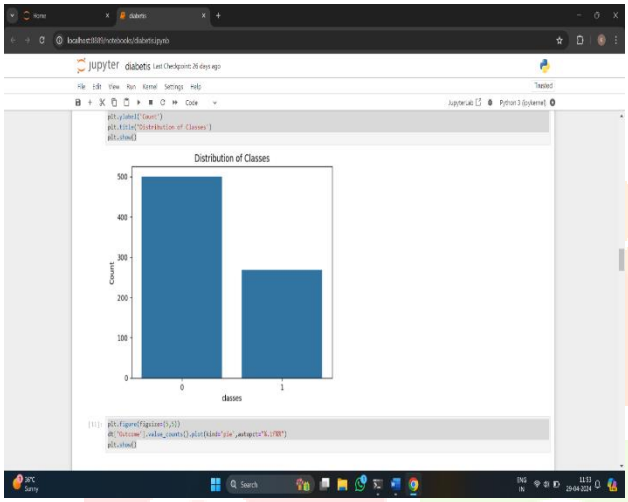


Figure:5.7 Distribution of Classes

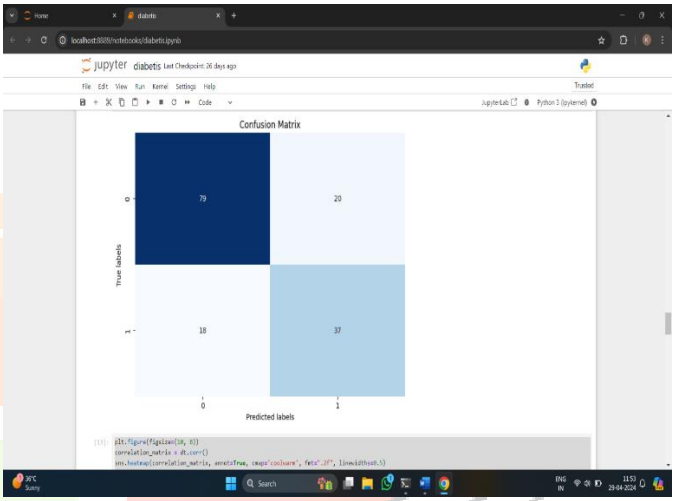


Figure: 5.8 Confusion Matrix

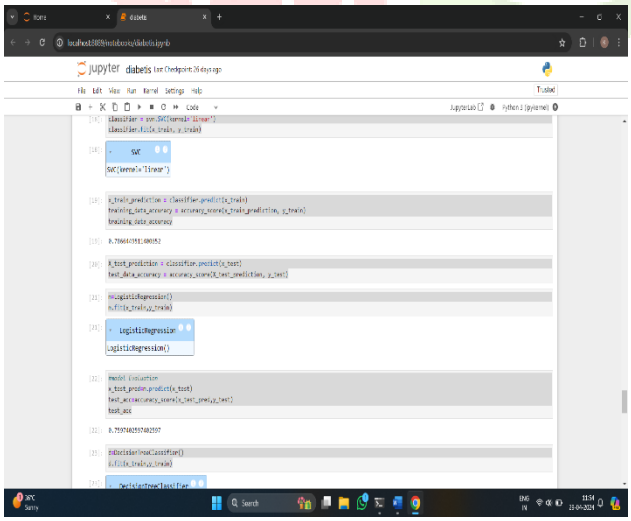


Figure 5.9: Accuracy of Svm and Logistic Regression

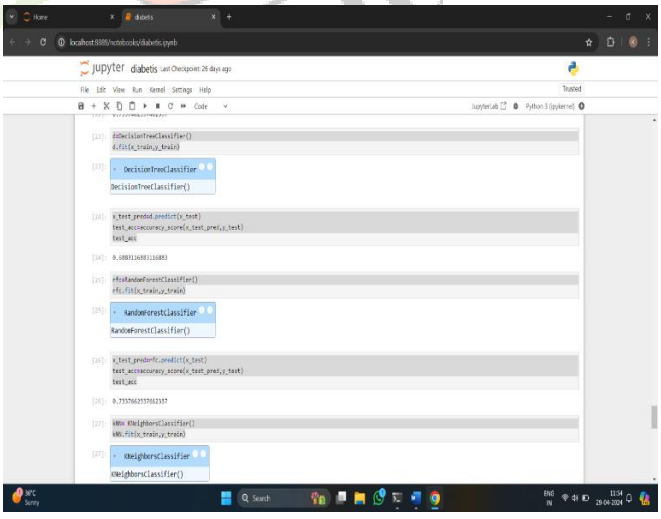


Figure 5.9: Accuracy of Svm and Logistic Regression

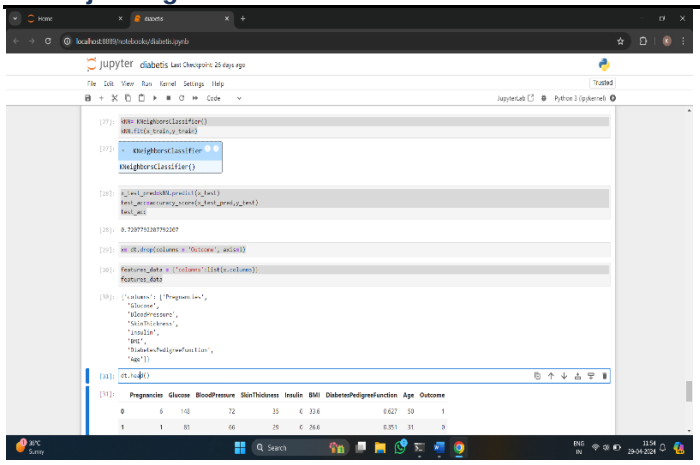


Figure 5.11: Accuracy of KNN

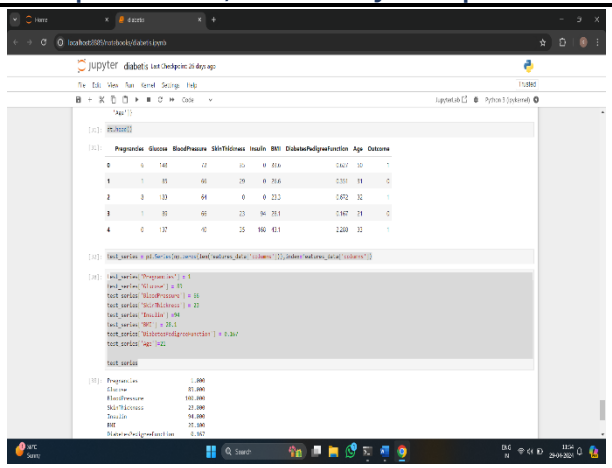


Figure 5.13: Test Series

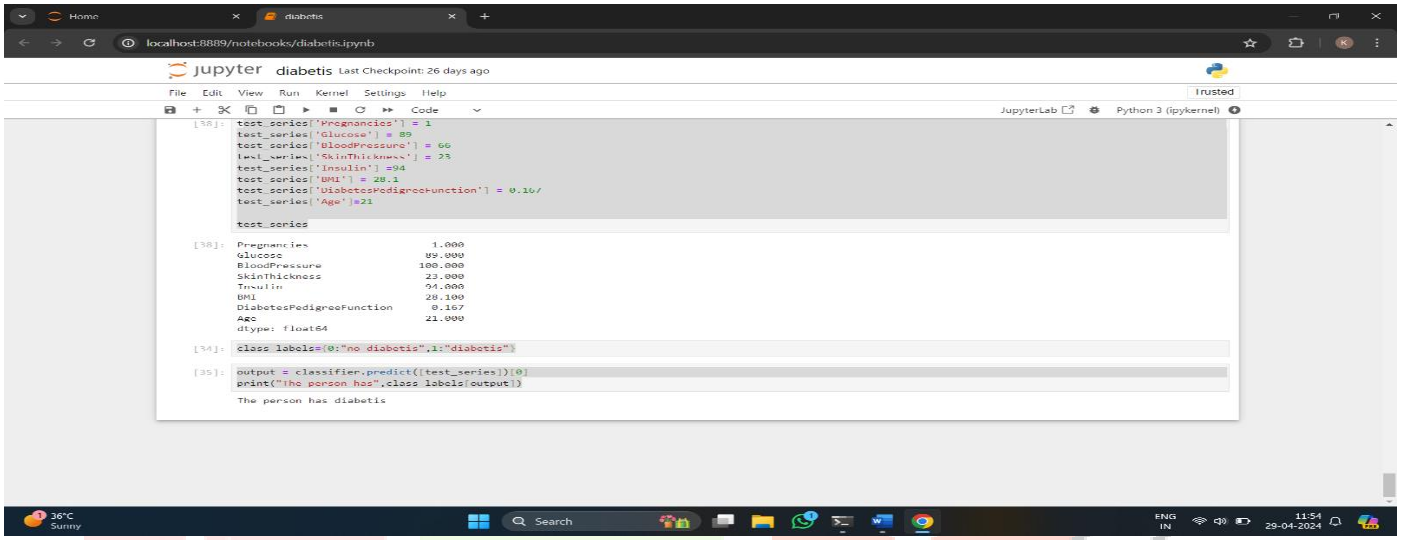


Figure 5.14: Out Come (Out Put)

VI. CONCLUSION

Goal is to build a model using supervised learning methods that could help assist doctors in the detection of diabetes to improve the quality of patient's lives. The paper presented multiple techniques that were used to train multiple models, support vector machine (svm) achieved the highest accuracy of 78%.

VII. FUTURE SCOPE

- Personalized Predictive Models: Future advancements in machine learning could lead to the development of more personalized predictive models for diabetes detection. These models could take into account a wider range of data sources, including genetic factors, lifestyle behaviors, and environmental influences, to provide more accurate risk assessments and early detection.
- Integration of Wearable Technology : With the increasing popularity and capabilities of wearable devices, such as continuous glucose monitors and fitness trackers, there is potential to integrate data from these devices into machine learning algorithms for diabetes detection. This could enable real-time monitoring and early detection of fluctuations in blood sugar levels, allowing for timely interventions and improved management of the disease.
- Explainable AI and Clinical Adoption: As machine learning algorithms become more complex, there is a growing need for explainable AI techniques to ensure transparency and trust in the decision-making process, especially in healthcare settings. Future research could focus on developing interpretable models that not only provide accurate predictions but also explain the reasoning behind their decisions, facilitating their adoption by clinicians and patients for diabetes detection and management.

VIII. REFERENCE

1. Swapna, G., R. Vinaya kumar, and K. P. Soman. "Diabetes detection using deep learning algorithms." *ICT express* 4.4 (2018): 243-246.
2. Abdulhadi, Nour, and Amjed Al-Mousa. "Diabetes detection using machine learning classification methods." *2021 international conference on information technology (ICIT)*. IEEE, 2021.
3. Gujral, Sakshi. "Early diabetes detection using machine learning: a review." *Int. J. Innov. Res. Sci. Technol* 3.10 (2017): 57-62.
4. Mujumdar, Aishwarya, and Vb Vaidehi. "Diabetes prediction using machine learning algorithms." *Procedia Computer Science* 165 (2019): 292-299.
5. Farajollahi, Boshra, et al. "Diabetes diagnosis using machine learning." *Frontiers in Health Informatics* 10.1 (2021): 65.
6. Kamble, Ms TP, and S. T. Patil. "Diabetes detection using deep learning approach." *International Journal for Innovative Research in Science & Technology* 2.12 (2016): 342-349.
7. Chaki, Jyotismita, et al. "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review." *Journal of King Saud University-Computer and Information Sciences* 34.6 (2022): 3204-3225.
8. Khaleel, Fayroza Alaa, and Abbas M. Al-Bakry. "Diagnosis of diabetes using machine learning algorithms." *Materials Today: Proceedings* 80 (2023): 3200-3203.
9. Warke, Mitesh, et al. "Diabetes diagnosis using machine learning algorithms." *Diabetes* 6.03 (2019): 1470-1476.
10. Theerthagiri, Prasannavenkatesan, A. Usha Ruby, and J. Vidya. "Diagnosis and classification of the diabetes using machine learning algorithms." *SN Computer Science* 4.1 (2022): 72.